Effective Image Super Resolution via Hierarchical Convolutional Neural Network

Bangli Liu, and Djamel Ait-Boudaoud*

jlliubangli@gmail.com, dab@port.ac.uk

University of Portsmouth, UK

Abstract

An extensive amount of research work has been carried out in image super-resolution using convolutional neural networks. The focus of a significant number of reported approaches is primarily on either increasing the depth or the width of the networks to achieve improvements in performance. This paper proposes a novel hierarchical convolutional neural network (HCNN) for effective image super-resolution by learning features from different levels. More specifically, the proposed framework implements a 3-step hierarchical process, which consists of an edge extraction branch, an edge reinforcement branch, and an image reconstruction branch. Informative edges in an image are extracted and enhanced in the edge extraction and reinforcement branch, which are used as a guidance in the image reconstruction branch. Experimental results on several public datasets demonstrate that the proposed framework can restore high-frequency edges information and achieve superior performances over the state-of-the-art methods. Moreover, through

*Corresponding author.
Email address: dab@port.ac.uk

a case study in facial expression recognition, we show that the enhanced images are beneficial in improving the recognition performance, paving the way for more practical applications.
Keywords:
Image super-resolution, Edges extraction, Deep learning, Facial expression recognition

---

1. Introduction

The task of image super-resolution (SR) is to restore high-resolution (HR) images from low-resolution (LR) images caused by down-sampling or blurring. It has wide applications, ranging from surveillance video recovery [1], medical imaging [2], object recognition [3], to satellite imaging [4]. It can also improve the performance of various vision related tasks such as face or facial expression recognition [5, 6, 7]. However, image SR is a highly ill-posed problem since there are multiple solutions for transforming LR images. To address this problem, many SR approaches have been developed, which can be broadly divided into interpolation-based and learning-based. Interpolation-based methods, such as bilinear, bicubic, and spline, tend to use a single kernel to estimate the unknown pixels in HR images by using the features of their neighbourhood pixels. The interpolation methods are frequently used due to their computational simplicity, however, their interpolated results based on a sole kernel are prone to blurriness, especially with large upscale factors.

Learning-based methods such as random forest [8], sparse coding [9, 10], neighbor embedding [11] are commonly used to model a mapping between LR

and HR patches. Motivated by the success of deep neural networks in image classification related tasks, many approaches have been proposed to learn the mapping between LR images and HR images via different structures of networks. For example, Dong *et al* [12] proposed a deep convolutional neural network for image super-resolution (SRCNN), where an end-to-end mapping between LR and HR images was learnt in a lightweight structure with three layers. SRCNN successfully introduced the deep learning technique into the image SR problem and many different neural networks have been proposed thereafter. However, most of these concentrated on directly learning a one-step mapping function from LR images to HR images, which usually fail to effectively recover high-frequency information, such as details of edges.

A HR image is made up of the low-frequency and high-frequency components. Compared to low-frequency information, high-frequency information is more easily corruptible resulting in noticeable image degradations. The details of edges and texture information are stored in the high-frequency component. Thus, restoring these details is an essential step to improve the quality of images. Observing the fact that the most perceptually salient features are represented by edges in an image and given the fact that it is much easier to extract the edges than texture information in degraded images, it is conjectured that detecting edges as prior knowledge is highly likely to improve the construction of HR images. Recently, Yang *et al* [13] extracted edges from LR image input by using a hand-crafted edge detector and fed them into the network together with the raw LR image. Although the hand-crafted edges can benefit the process of SR in some cases, they are normally computationally expensive and are not applicable for all types of images. To

better exert the beneficial guidance of edge information on image SR, this paper proposes a novel hierarchical convolutional neural network to perform image SR with suitably modeled edge prior.

Although the image sensing technology has been greatly improved in recent decades, it still remains a great challenge in capturing high-resolution facial images in some scenarios such as video surveillance and human-computer interaction. The low-resolution of facial images is likely to be one of the factors that affects the performance of existing facial expression recognition algorithms. On the other hand, it is believed that the SR technique with the improved quality of images has the potential to improve the performance of many vision related tasks. Thus, this paper further explores the possibility of using the SR technique to improve facial expression recognition.

The contributions of this paper are summarized as follows:

1. Unlike existing research which focuses on increasing the width or the depth of networks, this paper develops a novel hierarchical convolution neural network to adaptively extract and fuse image features at different levels. The edge information extracted by the shallow network branches is shown to deliver outstanding performance.

2. Extensive experimental evaluation on several publicly available datasets has demonstrated the superior performance of the proposed HCNN over the existing methods.

3. A case study in facial expression recognition proves that the enhanced images are beneficial in improving the recognition performance, paving the way for more practical applications.

The remainder of this paper is organized as follows: Section 2 reviews re-

lated work of deep learning based image super-resolution. Section 3 describes the proposed hierarchical convolutional neural networks for image SR. Section 4 presents experimental results as well as a comparison with the leading methods on three public datasets. This section further discusses an application of the proposed SR algorithm in facial expression recognition. Section 5 concludes the paper and discusses potential directions of the future work.

2. Related work

With the development of deep learning techniques, different deep neural networks have been proposed for the image super-resolution task. Dong *et al* [12] proposed SRCNN to learn an end-to-end mapping between low- and high-resolution images, where patch representation, non-linear mapping between a low-resolution patch and a high-resolution patch, and reconstruction were formulated as a convolutional operation. Later, they further proposed FSRCNN [14] which improves the SRCNN by introducing shrinking, expanding, and deconvolution layers. In these methods, the LR image is upscaled to the size of the desired HR images before input to the neural networks, indicating that the SR operation is performed in HR space. Shi *et al* [15] argued that this is sub-optimal and adds computational complexity. Thus, they introduced an efficient sub-pixel convolution layer which learns an array of upscaling filters to upscale the final LR feature maps into the HR output. This could reduce the computational and memory complexity by increasing the resolution from LR to HR only at the end of the network.

Instead of magnifing LR images to HR images at the first or final layer, the progressively upsampling strategy is proposed in [16] [17]. For example, Yang

et al [16] developed a gradual upsampling network, which uses a gradual process to simplify the difficult direct SR problem to an easier multistep upsampling task with a very small magnification factor in each step. Lai *et al* [17] proposed a pyramid super-resolution network to progressively reconstruct HR images in a coarse-to-fine fashion. Their proposed network progressively reconstructed the sub-band residuals of high-resolution images at multiple pyramid levels instead of the bicubic interpolation for pre-processing, thus it directly extracts features from the low-resolution input space and thereby entails low computational loads.

Inspired by VGG-net used for ImageNet classification, some researchers adopted very deep convolution networks to deal with the image super resolution problem [18, 19, 20, 21]. Kim *et al* [18] proposed a Deeply-Recursive Convolutional Network (DRCN) by introducing a very deep recursive layer via a chain structure with up to 16 recursions. They further proposed recursive-supervision and skip-connection to solve the gradient explosion problem. Kim *et al* [19] proposed a very deep convolution network (VDSR) including 20 weight layers. The slow convergence issue is tackled by using very high learning rates (10^{-1}). Compared to SRCNN which relies on the context of small image regions, they efficiently exploited context information over large image regions. The network learned the residual image instead of HR image directly and the original LR image is interpolated before it goes through networks. Therefore, the final output is the learned residual image plus the interpolated LR image. Tai *et al* [21] proposed deep recursive residual networks up to 52 convolutional layers. Both global and local residual learning were adopted to mitigate the difficulty of training. Recursive learning is used

to control the model parameters while increasing the depth.

Zhang *et al* [22] proposed a dimensionality stretching strategy that enables a single convolutional super-resolution network to take two key factors of the single image SR degradation process, i.e., blur kernel and noise level, as input. Consequently, the proposed super-resolver can handle multiple and even spatially variant degradations, which significantly improves its practicability. Since real LR images rarely obey the assumption of bicubic downscaling, Shocher *et al* [23] proposed to exploit the internal recurrence of information inside a specific image rather than relying on prior training, where training data are normally bicubicly downsampled from their HR images. A small image-specific CNN at test time is trained using the examples solely from the input image itself.

Yang *et al* [13] argued that the LR image and its edge map can jointly infer the sharp edge details of the HR image. Thus, they proposed a recurrent residual learning which has the edge-preserving capacity to effectively recover the difference between LR images and HR images. The by-pass connections across multiple layers were constructed to speed up the training convergence rate. Chen *et al* [24] believed that the specific facial prior knowledge could be helpful for better super-resolving face images. Thus, the geometry prior, i.e. facial landmark heatmaps and parsing maps were used in their face super-resolution network.

3. Proposed Method

3.1. Motivation

A HR image Y_{HR} could be decomposed into two components:

$$Y_{HR} = Y_{LF} + Y_{HF} \tag{1}$$

where Y_{LF} and Y_{HF} represent the low-frequency and high-frequency components, respectively. Y_{HF} contains subtle details of the image (such as edge and texture information), which are normally irregular and have smaller magnitude compared with Y_{LF}. This makes Y_{HF} easier to be corrupted in image degradation. To this end, in image super resolution tasks, restoring the details of the high-frequency component is an important step. However, most of the existing methods only depend on one branch for restoring the details of texture information. Considering that the deep convolution layers focus on

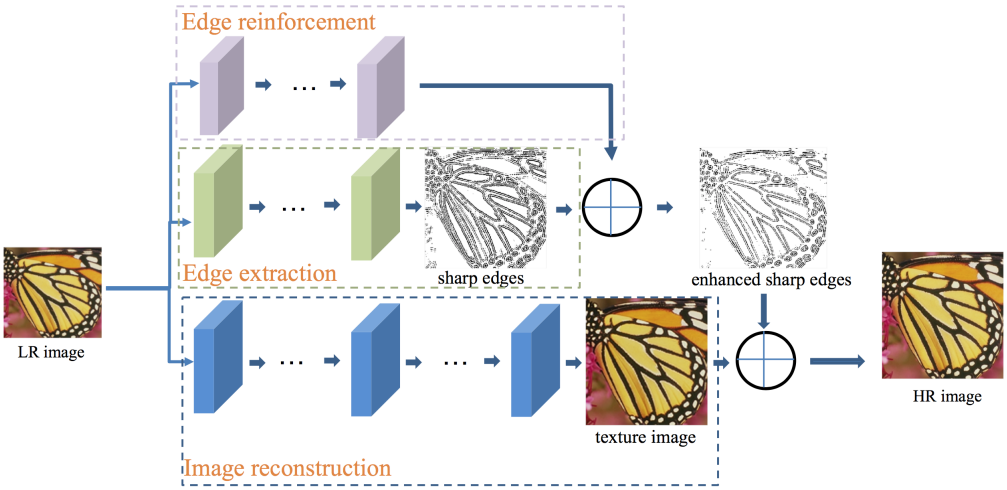

Fig. 1: Proposed hierarchical framework for image super resolution. It consists of edge extraction, edge reinforcement, and image reconstruction.

learning high-level image representations and might discard useful low-level information such as the image edge and other high-frequency signals, this paper proposes to enhance this information via stacking several shallow network structures, which is named as HCNN. As shown in Fig. 1, the HCNN consists of three functional networks for edge extraction, edge reinforcement, and image reconstruction.

Observing that sharp information is likely to be lost due to the deep convolutional operations, an edge extraction branch is introduced to retrieve the edges from LR images. This branch consists of 11 convolutional layers, where each layer contains 32 kernels with a size of 3*3. To further enhance the informative edge information, a much shallower edge reinforcement branch is introduced to assist the generation of sharp edges. This edge reinforcement branch is made up of 5 convolutional layers and each layer consists of 32 kernels with a size of 3*3. The texture information is restored via the image reconstruction branch which has 20 convolutional layers, where each layer contains 64 kernels with a size of 3*3. Among them, the first layer is designed to input the low-resolution image that needs to be improved and the last layer is to output the corresponding high-resolution image with given upscaling factors. The same rule applies to all the networks designed in this paper.

3.2. Sharp Edge Extraction

Since the edges contain the most perceptually salient features of an image, extracting edges from LR images should benefit the construction of HR images substantially. Thus, this paper proposes an edge extraction branch and an edge reinforcement branch to extract fine edge information. This kind of information then serves as a guidance for restoring HR images. The edge

extraction contains coarse edge extraction and sharp edge reinforcement. An edge extractor f_{edge} with parameters Θ_{edge} is learned in the edge extraction branch. It outputs an edge map with an input LR image Y_{LR}, which is represented as follows:

$$Y_{edge} = f_{edge}(Y_{LR}; \Theta_{edge}), \tag{2}$$

To highlight sharp edges in an image, an edge enhancement branch is further developed, whose output is a weight map $W_{\Theta_{rein}}$. The final sharp edges are calculated by the element-wise addition between Y_{edge} and $W_{\Theta_{rein}}$:

$$Y_{rein} = g(Y_{edge}; \Theta_{rein}) = W_{\Theta_{rein}} \oplus Y_{edge}, \tag{3}$$

where $\oplus$ is the element-wise addition.

Fig. 2 provides three examples of edges extracted by the edge extraction branch and the edge reinforcement branch, respectively. It can be seen that the edges enhanced by the reinforcement branch are sharper and more informative compared to the raw edges extracted by the extraction branch.

3.3. HR Image Reconstruction

To restore the texture features from the LR image Y_{LR}, a function f_{image} is modeled through the image reconstruction branch. The output $Y_{texture}$ is described as follows:

$$Y_{texture} = f_{image}(Y_{LR}; \Theta_{image}) \tag{4}$$

where Θ_{image} denotes the parameters of function f_{image}.

Fig. 2: Three examples of coarse edges and sharp edges extracted by the edge extraction branch and the edge reinforcement branch, respectively (upscale $*2$).

The final HR image is reconstructed using an element-wise sum of the texture image and sharp edge image:

$$\begin{aligned} Y_{HR} &= f_{image}(Y_{LR}; \Theta_{image}) \oplus g(f_{edge}(Y_{LR}; \Theta_{edge}); \Theta_{rein}) \\ &= Y_{texture} \oplus (W_{\Theta_{rein}} \oplus Y_{edge}), \end{aligned} \tag{5}$$

3.4. Loss Function

Given training data pairs $\{Y^i_{LR}, Y^i_{GT}\}^n_{i=1}$, the goal of the proposed neural network structure is to learn a model F with parameters Θ to predict the HR image $Y_{HR} = F(Y_{LR}; \Theta)$ by minimizing the error function. Thus, the objective function could be represented by the error average over the training data:

$$\min_{\Theta} \frac{1}{n} \sum_{i=1}^{n} ||F(Y^i_{LR}; \Theta) - Y^i_{GT})||^2_2 \tag{6}$$

where Y^i_{LR} and Y^i_{GT} are i^{th} LR image and ground truth image pair in the training data. n is number of image pairs. $F(Y^i_{LR}; \Theta)$ is the restored HR images using the proposed framework with parameters Θ.

4. Experiment

4.1. Experimental Settings

To conduct a fair comparison, instead of training our model on a large scale dataset, we train it using the training set used in the compared methods [19, 18, 25], which means that 91 images from [10] and 200 images from the training set of Berkeley segmentation dataset [26] are selected as training data. At training time, rotation and flip methods are applied for data augmentation. The learning rate begins with 10^{-3} and is divided by 10 at each $2 * 10^5$ training iterations. A weight decay of 0.001 and a momentum

of 0.9 are used in the training. The model is trained for $135 * 10^4$ iterations with a batch size of 32. At testing time, three publicly used datasets, namely, Set5 [11], Set14 [27], and BSD100 [28] are selected for evaluation. Set5 has 5 images: baby, bird, butterfly, head, woman. Set14 has 14 images: baboon, barbara, bridge, coastguard, comic, face, flowers, floreman, lenna, man, monarch, pepper, ppt3, zebra. BSD100 has 100 images with diverse natural scenes.

Three types of experiments have been conducted to comprehensively evaluate the performance of the proposed method. The first experiment compares the results of the HCNN with the state-of-the-art methods on three datasets (Set5, Set14, and BSD100). The second experiment aims to validate the advantage of the edge reinforcement branch. In this part, the performance of the HCNN and its baseline (HCNN without the edge reinforcement branch) is discussed. Finally, through a case study in facial expression recognition, the potential of the proposed SR technique in boosting the performance of recognition algorithms is explored.

4.2. Performance Comparison

This part compares the proposed method with the state-of-the-art methods in terms of the quality of SR images. Two criteria are used to evaluate the performance of the proposed method. PSNR(dB): Peak signal-to-noise ratio, and a higher PSNR generally indicates the reconstruction is of higher quality. SSIM: Structural similarity, measuring the similarity between two images. Bicubic interpolation and seven state-of-the-art methods: A+ [29], SelfEx [30], SRCNN [12], CSCN [31], VDSR [19], DRCN [18], DEGREE-3 [25], are selected for the performance comparison.

Table 1: Comparison of PSNR and SSIM on *Set5*, *Set14*, *BSD100* dataset for upscaling factors *2, *3, and *4.

Methods	-	Set5			Set14			BSD100		
		*2	*3	*4	*2	*3	*4	*2	*3	*4
Bicubic	PSNR	33.6	30.39	28.42	30.24	27.55	26.00	29.56	27.21	25.96
	SSIM	0.9299	0.8682	0.8104	0.8688	0.7742	0.7027	0.8431	0.7385	0.6675
A+	PSNR	36.54	32.58	30.28	32.28	29.13	27.32	31.21	28.29	26.82
[29]	SSIM	0.9544	0.9088	0.8603	0.9056	0.8188	0.7491	0.8863	0.7835	0.7087
SelfEx	PSNR	36.49	32.58	30.31	32.22	29.16	27.40	31.18	28.29	26.84
[30]	SSIM	0.9537	0.9093	0.8619	0.9034	0.8196	0.7518	0.8855	0.7840	0.7106
SRCNN	PSNR	36.66	32.75	30.48	32.45	29.30	27.50	31.36	28.41	26.90
[12]	SSIM	0.9542	0.9090	0.8628	0.9067	0.8215	0.7513	0.8879	0.7863	0.7101
CSCN	PSNR	36.88	33.10	30.86	32.50	29.42	27.64	31.40	28.50	27.03
[31]	SSIM	0.9547	0.9144	0.8732	0.9069	0.8238	0.7573	0.8884	0.7885	0.7161
VDSR	PSNR	37.53	33.66	31.35	33.03	29.77	28.01	31.90	28.82	27.29
[19]	SSIM	0.9591	0.9213	0.8838	0.9124	0.8314	0.7674	0.8960	0.7976	0.7251
DRCN	PSNR	37.63	33.82	31.53	33.04	29.76	28.02	31.85	28.80	27.23
[18]	SSIM	0.9588	0.9226	0.8854	0.9118	0.8312	0.7670	0.8942	0.7963	0.7233
DEGREE-3	PSNR	37.54	33.72	31.43	33.01	29.78	28.02	31.76	28.69	27.14
[25]	SSIM	0.9584	0.9204	0.8818	0.9118	0.8317	0.7646	0.8939	0.7937	0.7200
Baseline	PSNR	34.74	32.34	30.51	31.59	29.01	27.51	30.62	28.16	26.81
	SSIM	0.9386	0.9023	0.8644	0.9063	0.8258	0.7618	0.8926	0.7948	0.7223
HCNN	PSNR	37.62	33.77	31.39	33.03	29.79	28.04	31.91	28.84	27.29
	SSIM	0.9594	0.9230	0.8849	0.9127	0.8318	0.7674	0.8965	0.7985	0.7260

Table 1 lists the quantitative comparisons with upscale factors $*2, *3$ and $*4$ on three datasets. The underlined results represent the highest achieved performance in each column. It can be seen that the HCNN outperforms other listed methods in most cases. It performs the best on the BSD100 dataset with the highest PSNR and SSIM for all three upscale factors. Compared to DEGREE-3 which also employed edges extraction using a hand-crafted method as prior knowledge, the proposed method learning the informative edges via the edge extraction and the edge reinforcement branch

achieves the largest gain of 0.15dB (PSNR) and 0.0048 (SSIM) with the upscale factor $*3$. In addition, the superior performance achieved by HCNN over its baseline method has clearly demonstrated the benefit of adding the edge reinforcement branch.

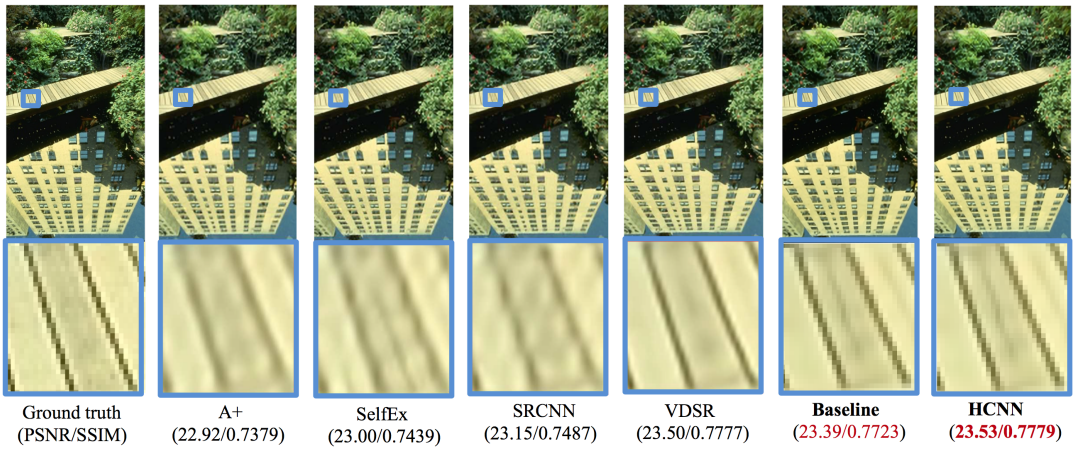

Fig. 3: SR results of *148026* in the BSD100 dataset (upscale $*3$). **Baseline** and **HCNN** are the results achieved by the proposed methods without and with the edge reinforcement branch respectively.

Fig. 3 shows the visual results of different methods on the image '148026' from the BSD100 dataset with the upscale factor $*3$. It can be seen that both proposed methods (both HCNN and its baseline method) achieve good quality HR image reconstruction. Specifically, compared to the blurry results (e.g., edges) achieved by methods A+, SelfEx, and SRCNN, the much clearer results (e.g., finer edges) can be obtained using the proposed methods with large margin of PSNR and SSIM. Although VDSR has high PSNR and SSIM values, the details in the image are oversmoothed. By contrast, benefiting from the valuable guidance from informative edges, the proposed HCNN

method delivers the highest PSNR and SSIM values.

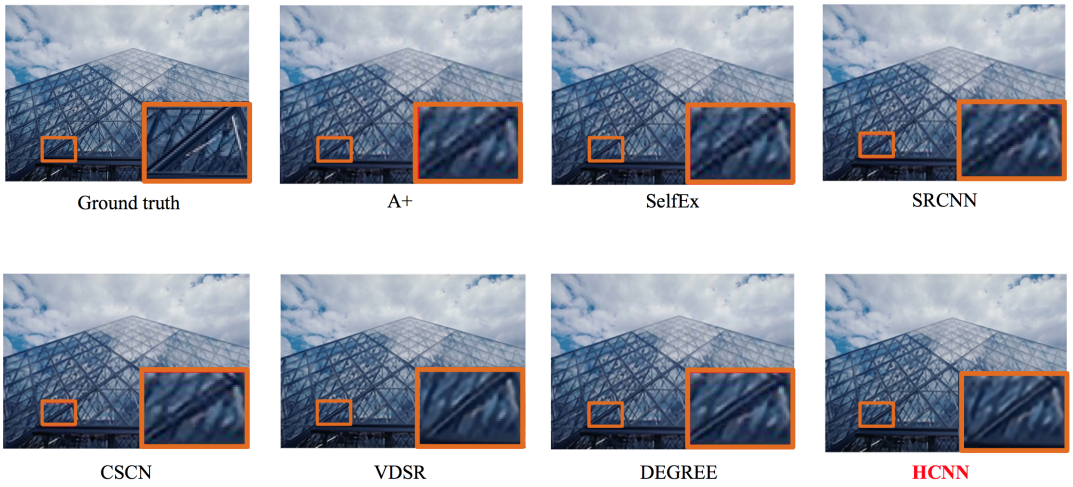

Fig. 4: SR results of *223061* from the BSD100 dataset (upscale $*3$). The result images of A+, SelfEx, SRCNN, CSCN and DEGREE are reproduced from [25] and the result of VDSR is produced using the open-sourced code.

Fig. 4 shows the restored results ($*3$) of the image "223061" from the BSD100 dataset achieved by different approaches. Methods like A+, SelfEx, SRCNN, and CSCN generate the results with blurry edges and textures. Although the result from VDSR improves the clarity of edges, it is over-smoothed. Owing to the informative edges learned from the edge enhancement branch, our method achieves sharp edges as well as fine textures. It also outperforms DEGREE which used hand-crafted extracted edges as prior knowledge.

4.3. Evaluation of Edge Reinforcement

To demonstrate the effectiveness of integrating an edge enhancement branch in the proposed method, the detailed super-resolution results on im-

Table 2: Super resolution quality of the HCNN method (with edge enhancement) and its baseline method (without edge enhancement) on Set5 dataset for upscaling factors *2, *3, and *4 (PSNR/SSIM).

Image	Baseline	HCNN	Baseline	HCNN	Baseline	HCNN
	Scale *2		Scale *3		Scale *4	
Baby	35.50/0.9615	38.79/0.9673	33.61/0.9218	35.45/0.9277	32.19/0.8829	33.44/0.8889
Bird	37.06/0.9539	42.60/0.9895	34.60/0.9299	36.88/0.9665	32.36/0.8921	32.54/0.9294
Butterfly	33.11/0.9754	34.70/0.9762	29.44/0.9431	30.05/0.9440	26.88/0.9074	27.21/0.9082
Head	33.99/0.8386	35.96/0.8905	32.36/0.7842	34.00/0.8346	31.73/0.7427	32.78/0.7916
Woman	34.03/0.9637	36.06/0.9734	31.42/0.9325	32.45/0.9422	29.36/0.8969	29.96/0.9065
Average	**34.74/0.9386**	**37.62/0.9594**	**32.34/0.9023**	**33.77/0.9230**	**30.51/0.8644**	**31.39/0.8849**

Table 3: Super resolution quality of the HCNN method (with edge enhancement) and its baseline method (without edge enhancement) on Set14 dataset for upscaling factors *2, *3, and *4 (PSNR/SSIM).

Image	Baseline	HCNN	Baseline	HCNN	Baseline	HCNN
	Scale *2		Scale *3		Scale *4	
baboon	25.72/0.7791	25.97/0.7801	23.64/0.6214	23.79/0.6223	22.70/0.5149	22.82/0.5156
barbara	27.69/0.8733	28.08/0.8743	25.90/0.7790	26.17/0.7800	25.73/0.7434	25.98/0.7444
bridge	27.67/0.8582	28.06/0.8589	25.15/0.7224	25.37/0.7229	23.78/0.6116	23.94/0.6121
coastguard	30.22/0.8535	30.97/0.8546	26.99/0.6715	27.33/0.6725	25.98/0.5685	26.26/0.5693
comic	28.90/0.9321	29.44/0.9328	24.91/0.8121	25.13/0.8128	22.88/0.6933	23.02/0.6939
face	33.98/0.8393	35.94/0.8906	32.63/0.7839	33.98/0.8335	31.66/0.7412	32.71/0.7895
flowers	32.90/0.9424	34.37/0.9469	29.42/0.8686	30.02/0.8729	27.36/0.7972	27.73/0.8014
foreman	34.64/0.9733	37.13/0.9737	33.330.9484	35.00/0.9488	32.13/0.9250	33.33/0.9254
lenna	34.62/0.9322	37.08/0.9330	32.62/0.8912	34.01/0.8920	31.08/0.8551	32.00/0.8558
man	30.67/0.8730	31.44/0.8964	28.37/0.7918	28.81/0.8155	26.88/0.7208	27.19/0.7442
monarch	35.86/0.9808	39.54/0.9816	33.12/0.9603	34.72/0.9612	30.75/0.9342	31.16/0.9350
pepper	34.95/0.9236	37.44/0.9248	33.62/0.8969	35.33/0.8979	32.49/0.8752	33.73/0.8760
ppt3	31.66/0.9867	32.66/0.9833	27.30/0.9585	27.66/0.9522	25.30/0.9210	25.52/0.9148
zebra	32.84/0.9402	34.25/0.9462	29.09/0.8550	29.63/0.8598	26.42/0.7641	26.69/0.7661
Average	**31.59/0.9063**	**33.03/0.9127**	**29.01/0.8258**	**29.78/0.8317**	**27.51/0.7618**	**28.04/0.7674**

ages from Set5 and Set14 are compared in Table 2 and 3 respectively. Compared to the baseline method without the edge reinforcement branch, the

proposed HCNN consistently achieves better performance in improving the image resolution on two datasets, which is proven by higher PSNR/SSIM on each image for three upscaling factors (*2, *3, and *4). For example, the PSNR on the image 'Bird' in Set5 is improved by 5.54dB by the HCNN, because of the extracted fine edge information serving as prior knowledge.

4.4. Application Example in Facial Expression Recognition

The low-resolution of face images is likely to be one of the reasons that decreases the performance of existing facial expression recognition algorithms in some public scenarios. By improving the quality of face images, this section aims to test the improvement of the proposed SR technique in the facial expression recognition task.

In this part, the raw low-quality face images in the FER-2013 dataset [32] are taken as input. FER-2013 dataset was collected by Pierre Luc Carrier and Aaron Courville for facial expression recognition research. It is created using the Google image search API to search for facial images that match a set of 184 emotion-related keywords such as "blissful" and "enraged". It includes six types of facial expressions: neutral, happy, sadness, surprise, anger, disgust, and fear. 28,709 gray images are used for training and 7178 images are used for validation. All searched images were resized to 48*48 pixels and converted to grayscale.

Most of them are captured in wild settings with low-resolution, which makes facial expression recognition challenging. The human accuracy on this dataset was just $68 \pm 5\%$. Upgrading the quality of greyscale images could improve the accuracy of facial expression recognition, however, the reasons causing low-resolution images are numerous, which makes it more difficult to

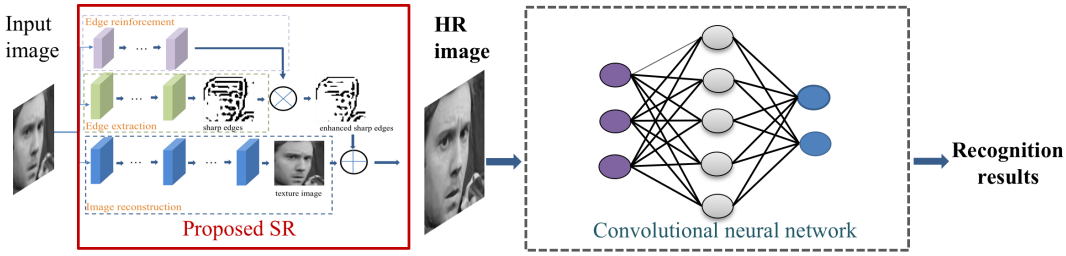

Fig. 5: Framework of the HCNN based facial expression recognition.

restore the HR images from these images compared to those downsampled by known methods. We use the proposed HCNN method to generate the HR facial expression images, and then apply convolutional neural network for expression recognition.

Fig. 6: The improved quality of images for facial expression recognition with upscaling factor $*2$.

Fig. 5 shows the process of the proposed end-to-end facial expression recognition. The HR images restored from input images via the HCNN are used for facial expression recognition in the following neural networks.

Fig. 6 gives some examples of high-quality images on the FER-2013 dataset restored from corresponding low-quality images via the proposed HCNN. The finer facial contours are reconstructed with visually plausible boundaries and edges, which makes facial expression easily recognizable.

Table 4 compares the performance of different methods. Here, Alexnet is taken as an example for HR facial expression recognition, but we believe that such improvement is applicable to other neural networks. It can be seen that the proposed HCNN+Alexnet using the SR technique as data pre-processing achieves the better performance over other existing methods with the highest accuracy of 69.1.

Table 4: Recognition accuracies (%) on the FER-2013 dataset.

[33]	66.4
[34]	67.79
FSN[35]	67.6
Alexnet[36]	61.1
HCNN+Alexnet	**69.1**

5. Conclusion

This paper proposed the HCNN framework which hierarchically assembles shallow CNNs with deep CNNs for effective image super-resolution. The edge information extracted in shallow CNNs outlines high-frequency features of the input image. It serves as supplementary information for deep CNNs to reconstruct high-resolution images. The HCNN was shown to achieve a great performance improvement on three commonly used datasets. Moreover,

it has been demonstrated that an improvement in classification accuracy can be achieved by integrating the proposed method with the existing facial expression recognition algorithm. Future direction will focus on exploring the potential application of image SR in areas where an improvement of image quality is needed, such as motion understanding for public video surveillance [37] and human gaze estimation for human-computer interaction [38].

References

[1] L. Zhang, H. Zhang, H. Shen, and P. Li, "A super-resolution reconstruction algorithm for surveillance images," *Signal Processing*, vol. 90, no. 3, pp. 848–859, 2010.

[2] H. Greenspan, "Super-resolution in medical imaging," *The Computer Journal*, vol. 52, no. 1, pp. 43–63, 2008.

[3] M. S. Sajjadi, B. Scholkopf, and M. Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4491–4500.

[4] M. W. Thornton, P. M. Atkinson, and D. Holland, "Sub-pixel mapping of rural land cover objects from fine spatial resolution satellite sensor imagery using super-resolution pixel-swapping," *International Journal of Remote Sensing*, vol. 27, no. 3, pp. 473–491, 2006.

[5] B. K. Gunturk, A. U. Batur, Y. Altunbasak, M. H. Hayes, and R. M. Mersereau, "Eigenface-domain super-resolution for face recognition,"

IEEE transactions on image processing, vol. 12, no. 5, pp. 597–606, 2003.

[6] W. W. Zou and P. C. Yuen, "Very low resolution face recognition problem," *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 327–340, 2012.

[7] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep cascaded bi-network for face hallucination," in *European Conference on Computer Vision*. Springer, 2016, pp. 614–630.

[8] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3791–3799.

[9] R. Timofte, V. De Smet, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1920–1927.

[10] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE transactions on image processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[11] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," in *BMVC*, 2012.

[12] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[13] W. Yang, J. Feng, J. Yang, F. Zhao, J. Liu, Z. Guo, and S. Yan, "Deep Edge Guided Recurrent Residual Learning for Image Super-Resolution," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5895–5907, 2017.

[14] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9906 LNCS, pp. 391–407, 2016.

[15] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1874–1883, 2016.

[16] Y. Zhao, G. Li, W. Xie, W. Jia, H. Min, and X. Liu, "Gun: Gradual upsampling network for single image super-resolution," *IEEE Access*, vol. 6, pp. 39 363–39 374, 2018.

[17] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Fast and accurate image super-resolution with deep laplacian pyramid networks," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[18] J. Kim, J. Kwon Lee, and K. Mu Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1637–1645.

[19] ——, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654.

[20] X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Advances in neural information processing systems*, 2016, pp. 2802–2810.

[21] Y. Tai, J. Yang, and X. Liu, "Image Super-Resolution via Deep Recursive Residual Network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, no. July, pp. 2790–2798, 2017.

[22] K. Zhang, W. Zuo, and L. Zhang, "Learning a single convolutional super-resolution network for multiple degradations," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 6, 2018.

[23] A. Shocher, N. Cohen, and M. Irani, ""zero-shot" super-resolution using deep internal learning," in *Conference on computer vision and pattern recognition (CVPR)*, 2018.

[24] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "Fsrnet: End-to-end learning face super-resolution with facial priors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2492–2501.

[25] W. Yang, J. Feng, J. Yang, F. Zhao, J. Liu, Z. Guo, and S. Yan, "Deep edge guided recurrent residual learning for image super-resolution," *IEEE Transactions on Image Processing*, vol. 26, no. 12, pp. 5895–5907, 2017.

[26] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5, pp. 898–916, 2011.

[27] R. Zeyde, M. Protter, and M. Elad, "Technical Report CS-2010-12 - 2010 On Single Image Scale-Up using Sparse-Representation," vol. 1, no. 1, pp. 1–10, 2010.

[28] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluation segmentation algorithms and measuring ecological statistics," *Proc. 8th International Conference on Computer Vision*, vol. 2, no. July, pp. 416–423, 2001.

[29] R. Timofte, V. De Smet, and L. Van Gool, "A+: Adjusted anchored neighborhood regression for fast super-resolution," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 111–126.

[30] J.-b. Huang, A. Singh, and N. Ahuja, "Single Image Super-resolution from Transformed Self-Exemplars Supplementary material," *Computer Vision and Pattern Recognition*, pp. 5197–5206, 2015.

[31] Z. Wang, D. Liu, J. Yang, W. Han, and T. Huang, "Deep networks for image super-resolution with sparse prior," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 370–378.

[32] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing*, 2013, pp. 117–124.

[33] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, 2016, pp. 1–10.

[34] B. Sun, L. Li, G. Zhou, and J. He, "Facial expression recognition in the wild based on multimodal texture features," *Journal of Electronic Imaging*, vol. 25, no. 6, p. 061407, 2016.

[35] S. Zhao, H. Cai, H. Liu, J. Zhang, and S. Chen, "Feature selection mechanism in cnns for facial expression recognition," in *BMVCW*, 2018, pp. 1–12.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[37] B. Liu, H. Cai, Z. Ju, and H. Liu, "Rgb-d sensing based human action and interaction analysis: A survey," *Pattern Recognition*, vol. 94, pp. 1–12, 2019.

[38] H. Cai, Y. Fang, Z. Ju, C. Costescu, D. David, E. Billing, T. Ziemke, S. Thill, T. Belpaeme, B. Vanderborght *et al.*, "Sensing-enhanced ther-

apy system for assessing children with autism spectrum disorders: a feasibility study," *IEEE Sensors Journal*, vol. 19, no. 4, pp. 1508–1518, 2018.